Protein Science (1994), *3*:687–695. Cambridge University Press. Printed in the USA.

The three-dimensional profile method using residue preference as a continuous function of residue environment

KAM Y.J. ZHANG AND DAVID EISENBERG

UCLA-DOE Laboratory of Structural Biology and Molecular Medicine, Molecular Biology Institute, and Department of Chemistry and Biochemistry, University of California at Los Angeles, Los Angeles, California 90024-1570

(RECEIVED July 14, 1993; ACCEPTED February 10, 1994)

Abstract

In the 3-dimensional profile method, the compatibility of an amino acid sequence for a given protein structure is scored as the sum of the preferences of the residues for their environments in the 3D structure. In the original method (Bowie JU, Lüthy R, Eisenberg D, 1991, *Science 253*:164–170), residue environments were quantized into 18 discrete environmental classes. Here, amino acid residue preferences are expressed as a continuous function of environmental variables (residue area buried and fractional area buried by polar atoms). This continuous representation of residue preferences, expressed as a Fourier series, avoids the abrupt change of preference of residues in slightly different environments, as encountered in the original method with its 18 discrete environmental classes. When compared with the discrete 18-class representation of residue environments, this continuous 3D profile is found to be more sensitive in identifying sequences that fold into the profiled structure but share with it little sequence identity. The continuous 3D profile is also less sensitive to errors in environmental variables than is the discrete 3D profile. The continuous 3D profile can also be used to detect wrong folds or incorrectly modeled segments in an otherwise correct structure, as could the discrete 3D profile (Lüthy R, Bowie JU, Eisenberg D, 1992, *Nature 356*:83–85). Moreover, the progress of structure improvement during atomic refinement can also be monitored by examining the profile scores in a moving-window scan. Finally, by defining a functional form for profile scores, we open the way to profile atomic refinement in which an atomic structure adjusts to produce residue environments more compatible with the protein side chains.

Keywords: amino acid sequence analysis; homologous modeling; inverted protein folding; protein properties; structure prediction

The inverted protein folding problem, of finding which amino acid sequences fold into a known 3-dimensional structure, was addressed in the 3D profile method of Bowie et al. (1991) by finding sequences that are most compatible with the environments of the residues in the structure. In this method, the environment of each residue position within the folded protein is characterized on the basis of 3 properties: (1) the area of side chain that is buried by other protein atoms (referred to as "area buried" hereafter); (2) the fraction of side-chain area that is covered by polar atoms (referred to as "fraction polar" hereafter); and (3) the secondary structure. The secondary structure is classified in 3 states: helix, sheet, and coil, based on the main chain hydrogen-bonding pattern. The residues are first divided into 6 classes based on the area buried and fraction polar (see Fig. 4 of Bowie et al., 1991). Further subdividing these 6 classes by the 3 secondary structure states yields 18 environmental classes in total. In short, each position in a 3D structure can be assigned to 1 of the 18 environmental classes.

The preferences of the 20 amino acids for each of these 18 environmental classes, called "3D-1D scores," are derived from a set of well-refined protein structures, together with sets of sequences homologous to the sequence of the 3D structure. This classification of environments enables a protein structure to be coded by a sequence in an 18-letter alphabet, in which each letter represents the environmental class of a residue position. A 3D profile is constructed by associating with the environmental class of each position the 20 3D-1D scores of amino acids for this class. This original 3D profile, created from the representation of environments by 18 discrete classes, is referred to in the present paper as a "discrete 3D profile." Dynamic programming (Needleman & Wunsch, 1970; Smith & Waterman, 1981) was used to find the best match of a test sequence with the discrete

Reprint requests to: David Eisenberg, Molecular Biology Institute, University of California at Los Angeles, 405 Hilgard Avenue, Los Angeles, California 90024; e-mail: david@uclaue.mbi.ucla.edu.

3D profile. The method was tested on several families of proteins and was able to identify the structural similarity of proteins, some of which share no detectable sequence similarity (Bowie et al., 1991).

Lüthy et al. (1992) used the 3D profile method to assess the correctness of a protein model. They demonstrated that the comparison of the 3D profile, calculated from the model structure, with its own amino acid sequence can be used as an effective test of the accuracy of structure model. The 3D profiles of correct protein structures match their own sequences with high scores, whereas 3D profiles from incorrect protein models score poorly. An incorrectly modeled segment in an otherwise correct structure can be identified by a 3D profile window plot (Lüthy et al., 1992), in which the average profile score for a window of 21 residues is plotted against sequence number.

There are both strengths and weaknesses in representing a protein structure in terms of 18 discrete environmental classes. The major strength is that such a representation provides a means whereby a 3D structure can be represented by a string of letters, analogous to a protein amino acid sequence, but with no direct reference to the amino acids in any given sequence. Furthermore, the use of discrete classes is necessary for accumulating adequate statistics of 3D-1D scores when only a few structures and homologous sequences are available. The use of discrete classes also averages out noise in the data. But, the discrete representation of residue environments also has weaknesses. A major weakness is that the division into discrete classes means that because of sharp class boundaries, an infinitesimal change in the area buried or fraction polar of a residue can change its environmental class and thus can alter the residue preferences dramatically. Because of the finite precision of computer arithmetic, and because of the discrete algorithms used for area calculations, the environmental class of a residue computed with our program could depend even on the orientation of the molecule on the grid.

To overcome these shortcomings of the discrete environmental classes, we introduce here a continuous representation of the residue preferences as a function of the environmental variables area buried and fraction polar. The continuous function we chose is expressed as a Fourier series. The method used in deriving the coefficients in the Fourier series is described here, and the improved results of this continuous 3D profile are discussed.

Theory

The preference of residue i at secondary structure state j in environment (b, p) is defined as the information value $S_{ij}^0(b, p)$ (Fano, 1961):

$$S_{ij}^0(b, p) = \ln\left[\frac{P(i|j, b, p)}{P(i)}\right] = \ln\left[\frac{P(i, j, b, p)}{P(i)P(j, b, p)}\right], \quad (1)$$

where $P(i|j, b, p)$ is the conditional probability of finding residue i at secondary structure state j in environment (b, p) with area buried b and fractional polarity p; $P(i, j, b, p)$ is the joint probability of residue i, secondary structure state j, area buried b, and fractional polarity p; $P(i)$ is the a priori probability of residue i derived from amino acid compositions; $P(j, b, p)$ is the joint probability of secondary structure state j, buried area b, and fractional polarity p, where

$$P(i) = \sum_j \sum_{b,p} P(i, j, b, p) \quad (2)$$

$$P(j, b, p) = \sum_i P(i, j, b, p). \quad (3)$$

These quantities are evaluated as described by Bowie et al. (1991) by counting the number of residues of each type in each type of environment.

The residue preference at a given secondary structure state can be represented as a continuous function of its environments (b, p). We chose to use a 2-dimensional Fourier series as this continuous function.

$$S_{ij}(b, p) = \sum_{k,l} f_{kl} e^{-2\pi i(kb+lp)} \quad (4)$$

Notice that because the residue preference $S_{ij}(b, p)$ is a real function, the Fourier coefficient f_{kl} is therefore a complex Hermitian function. Hence, we need to sum only over the indices $k = [0, m]$ and $l = [-n, n]$, where m and n are the maximum orders of the indices.

If residue preferences $S_{ij}^0(b, p)$ are observed at some sampling points of (b, p) as described above, a set of Fourier coefficients f_{kl} for function $S_{ij}(b, p)$ that best represents the observed values $S_{ij}^0(b, p)$ can be evaluated by a least-squares minimization method.

Let the residual vector between the function $S_{ij}(b, p)$ and the observed data $S_{ij}^0(b, p)$ be represented by matrix notation,

$$\mathbf{r} = \{r_{ij}\} = \{S_{ij}(b, p) - S_{ij}^0(b, p)\}, \quad (5)$$

where $\{\ \}$ represents a matrix formed by a set of elements r_{ij}.

We need to minimize the residual $\mathbf{r}^T\mathbf{r}$:

$$\mathbf{r}^T\mathbf{r} = \{S_{ij}(b, p) - S_{ij}^0(b, p)\}^T\{S_{ij}(b, p) - S_{ij}^0(b, p)\}, \quad (6)$$

where the superscript T denotes transpose.

Let the partial derivative matrix be represented as

$$\mathbf{A} = \left\{\frac{\partial S_{ij}(b, p)}{\partial f_{kl}}\right\} = \{e^{-2\pi i(kb+lp)}\} \quad (7)$$

and the Fourier coefficient matrix be represented as

$$\mathbf{x} = \{f_{kl}\}, \quad (8)$$

where $k = [0, m]$ and $l = [-n, n]$.

Then we need to solve the following system of linear equations to get the Fourier coefficients, $\mathbf{x}$,

$$\mathbf{A}^T\mathbf{A}\mathbf{x} = \mathbf{A}^T\mathbf{b}, \quad (9)$$

where $\mathbf{b}$ is the negative of the residual vector:

$$\mathbf{b} = -\mathbf{r} = \{S_{ij}^0(b, p) - S_{ij}(b, p)\}. \quad (10)$$

Because the partial derivative matrix $\mathbf{A}$ is independent of $\mathbf{x}$, the coefficients $\mathbf{x}$ can be solved directly by matrix inversion:

$$\mathbf{x} = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{b}. \quad (11)$$

The inversion of matrix $(\mathbf{A}^T\mathbf{A})^{-1}$ is performed by Gauss-Jordan elimination method.

Therefore, given observed residue preferences $S_{ij}^0(b, p)$ at different values of the variables buried area b and fractional polarity p for residues i at a given secondary structure state j, we can evaluate the Fourier coefficients by a least-squares method, as described above.

We can use a smoothing factor, analogous to the Debye-Waller thermal factor (or B-factor) in X-ray crystallography, to average over a small area near the calculated area buried and fraction polar to reflect the accuracy of atomic positions in the model. This smoothing factor can be varied to select an optimal value for a given set of observations. The smoothing can be conveniently performed by multiplying the Fourier coefficients (Equation 4) by an exponential term,

$$S_{ij}(b, p) = \sum_{k,l} f_{kl} e^{-2\pi i(kb+lp)} e^{-B(k^2+l^2)}, \qquad (12)$$

where B is the smoothing factor.

Given these Fourier coefficients, the preferences of the 20 amino acid types for a position can be evaluated from the secondary structure and the values of the environmental variables area buried and fraction polar of that position. This representation of residue preferences by a continuous function serves 2 purposes. First, it creates a smooth surface from a discrete set of observed data. Second, it can be used to interpolate between data points.

A 3D profile (Bowie et al., 1991) can be created using this continuous representation of residue preferences in place of the original discrete representation. This new 3D profile, which has exactly the same matrix form as the original profile, is referred to as a continuous 3D profile.

Results

The continuous 3D-1D scoring surface

A database of 16 well-refined structures from the Protein Data Bank (Bernstein et al., 1977) and sets of highly homologous sequences aligned to the sequences of these structures (Lüthy et al., 1991) were used to generate the residue preferences in different environments. The secondary structures for all residues in the 3D structure were evaluated with the DSSP program (Kabsch & Sander, 1983). The buried surface area for each side chain and the fraction of side-chain area covered by polar atoms were calculated using the program ENVIRONMENTS-3D of Bowie et al. (1991).

We divided the area buried and fraction polar each into 32 equal bins for each of the 3 secondary structure classes of each of the 20 residues. The score value was from interpolating the 3D-1D scoring table of Bowie et al. (1991). Since the 3D-1D scoring surface is not a periodic function, the border regions are padded with the scores of their nearest neighbors in order to prevent aliasing. We then used a least-squares minimization method to evaluate the Fourier coefficients, as described in the Theory section. We chose to use a fourth-order Fourier series ($k = l = 4$ in Equation 12) to represent the residue preference surface. For each residue type at each secondary structure state, a set of 64-term Fourier coefficients was evaluated. Therefore, we have altogether 60 sets of 64-term Fourier coefficients to represent the preferences of all the residues in all secondary structure states.

Figure 1A shows a discrete 3D-1D scoring surface (Bowie et al., 1991) for tyrosine in the β-sheet secondary structure state. For comparison, Figure 1B shows the scoring surface for tyrosine in the β-sheet state as a continuous function of area buried and fraction polar, represented by a Fourier series. Both these scoring surfaces share the same general features of favorable and unfavorable regions of area buried and fraction polar. However, the smooth scoring surface in the Fourier series representation avoids abrupt changes of score when the area buried or fraction polar moves across class boundaries.

Tests of continuous 3D profiles

We tested continuous 3D profiles, created with the continuous representation of residue preferences as a function of environment, in searches for sequences that are compatible with the profiled structure.

3D compatibility searches with a continuous profile compared to that for a discrete profile for sperm whale myoglobin

In comparing continuous and discrete 3D profiles for the refined structure of sperm whale myoglobin (1MBO), we first determined the optimal value of the smoothing factor B by varying the value for B in Equation 12. The resulting continuous 3D profiles were used to score all protein sequences in a sequence database containing 59,091 nonidentical sequences. The relative effectiveness of different profiles is shown in Figure 2. In this figure, the most effective profiles are those with higher plots, as explained in the following. In a compatibility search with a given profile, each sequence receives a profile score for the optimal alignment of that sequence with the profile. These profile scores are then expressed as Z-scores, the number of standard deviations above the mean profile score normalized for sequence length. Then, sequences are ranked by their Z-scores and plotted as in Figure 2, where the number of globin sequences having that Z-score or a higher value ("Number of globins") is plotted on the ordinate as a function of the total number of sequences examined in decreasing Z-score. In this plot, a perfect profile would be represented as a line of slope 1 extending to the number of globin sequences in the database (691 in this example) and then would turn horizontal, because each additional sequence of lower Z-score would be a non-globin. For less than perfect profiles, some non-globins will have higher Z-scores than some globins, and the trace will fall below that for a perfect profile.

This relative effectiveness of various profiles can be seen more easily in the inset to Figure 2, which enlarges the region around 700 sequences examined. From this region, it appears that values between 0.0 and 0.1 for the smoothing factor B are the most discriminating for the recognition by the profile of its own and very similar sequences, and yet are still sensitive in identifying remotely related sequences. Figure 2, however, does not reveal the effect of the smoothing factor on the Z-scores assigned by the profile to the sequences. Figure 3 gives this information: it shows the number of globins correctly identified as a function of Z-score, comparing the discrete profile with continuous profiles having various smoothing factors. The "number of globins" is a cumulative number of all globins with Z-scores above the Z-score of the abscissa. Judging from Figure 3, the continuous

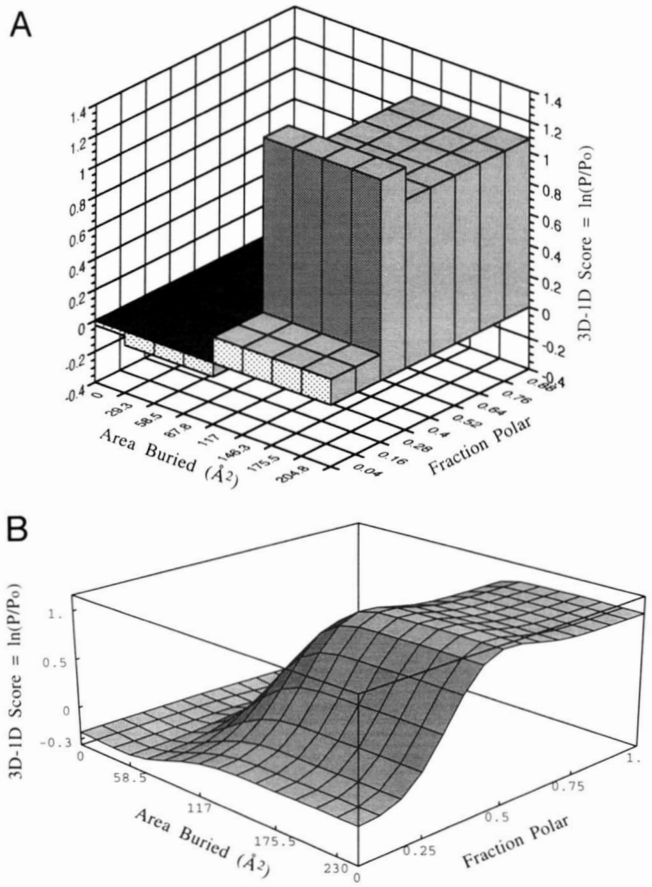

Fig. 1. A: 3D-1D preference scores for tyrosine in β-sheet for 6 discrete environmental classes. The preference is shown on the vertical axis as a function of the 2 environmental variables, area buried and fraction polar. **B:** 3D-1D preference for tyrosine in β-sheet as a continuous function of the environmental variables area buried and fraction polar. Notice that A and B have the same general shape, but that B lacks the discontinuous steps of A.

profile with $B = 0.1$ gives the best result. This profile, compared to the discrete profile and other continuous profiles, assigns higher Z-scores to its own sequence and closely related myoglobins and hemoglobins, and also higher Z-scores to even remotely related leghemoglobins (these are sequences with Z-scores around 8). The increased sensitivity of the continuous profile in the region of Z-score = 8 is of practical importance, because this is the typical Z-score for a distantly related sequence, which is compatible with the fold of the profiled protein.

3D compatibility searches with continuous profiles having various B-factors for dogfish muscle lactate dehydrogenase as compared with that of the discrete profile

To evaluate both discrete and continuous profiles with varying smoothing factors, we also compared the results of compatibility searches for a 3D profile prepared from the structure of lactate dehydrogenase (6LDH). Both malate dehydrogenase and alcohol dehydrogenase share the same dinucleotide binding motif (Rao & Rossmann, 1973) as lactate dehydrogenase. The sequence identity between malate dehydrogenases and lactate

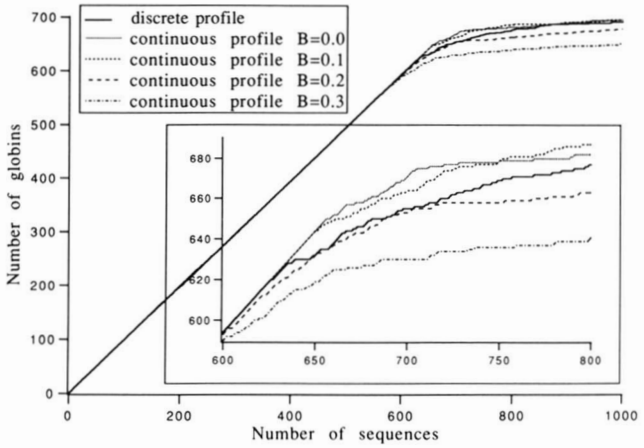

Fig. 2. Effectiveness of several profiles as indicated by the number of globins identified as a function of number of high-scoring sequences in a 3D profile compatibility search, using both discrete and continuous 3D profiles prepared from the structure 1MBO with various smoothing factors. A gap-opening penalty of 4.5 and gap-extension penalty of 0.05 were used for the compatibility searches. The inset is an enlargement of a region that is more important for comparison. A perfect selection is a straight line with slope equal to 1 that turns horizontal at the point corresponding to the number of globin sequences (691). Notice that the most effective B-value is 0.1.

dehydrogenases is about 23% on average. The sequence identity between alcohol dehydrogenases and lactate dehydrogenases is about 20% on average. Figure 4 shows the number of sequences and their average Z-scores identified for lactate, ma-

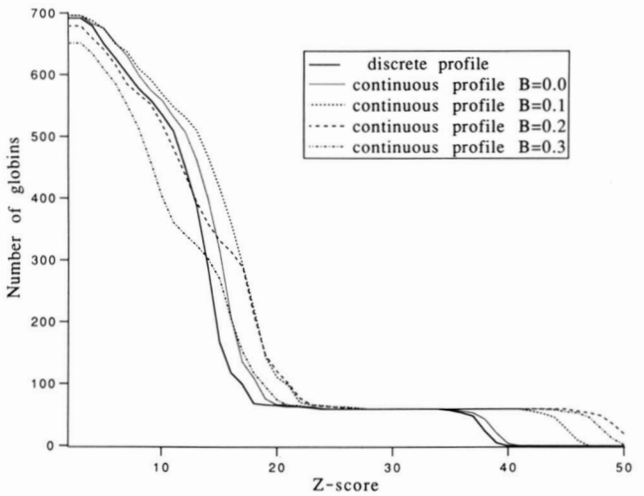

Fig. 3. Number of globin sequences assigned Z-scores above a threshold as a function of Z-scores in a compatibility search, using both discrete and continuous 3D profiles prepared from the structure 1MBO with various smoothing factors. Notice that the most effective B-value is 0.1. In the range of Z-scores 28–35, there are no globin sequences assigned. This corresponds roughly to the division of Z-scores between myoglobins and hemoglobins. Those sequences with Z-scores above 35 are mostly myoglobins, and those sequences with Z-scores between 10 and 28 are hemoglobins. Most of the leghemoglobins are assigned Z-scores between 5 and 10. The diagram shows that the continuous 3D profile does better than the discrete 3D profile both at high Z-score and at Z-score around 5–10. This low Z-score region is important for detecting distant sequences.

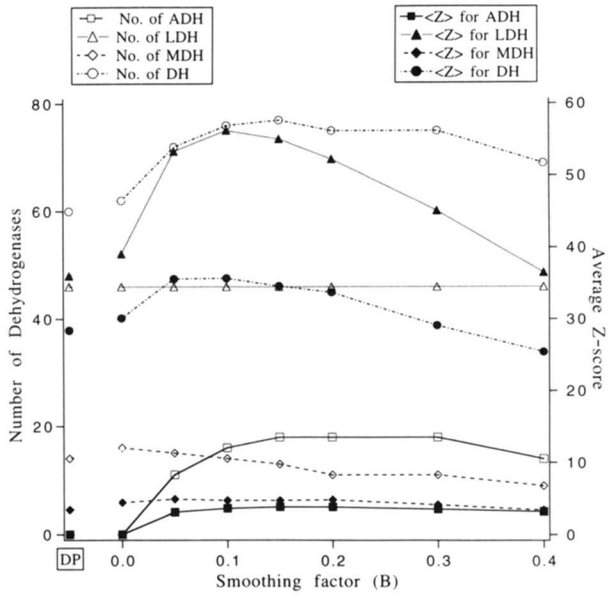

Fig. 4. Number of dehydrogenase sequences found to be compatible with continuous 3D profiles prepared from the structure 6LDH and their average Z-scores as a function of smoothing factors. The corresponding results for the discrete profile are also shown on the leftmost column for comparison. The ordinate is the smoothing factor. (Note: The leftmost column corresponds to the discrete profile.) The abscissa on the left is the number of dehydrogenases; the abscissa on the right is the average Z-score. ADH, alcohol dehydrogenase; LDH, lactate dehydrogenase; MDH, malate dehydrogenase; DH, dehydrogenase; DP, discrete 3D profile. Notice that the most effective B-value is 0.1.

late, and alcohol dehydrogenases from a compatibility search based on the 6LDH structure. Based on these results, the optimum value for the smoothing factor is about 0.1. Notice that the continuous profile identifies lactate dehydrogenase sequences with higher Z-scores than does the discrete profile (results shown at the left of the figure). Notice also that the continuous profile identifies more malate dehydrogenases and with overall higher Z-scores than does the discrete profile. Moreover, the continuous profile of 6LDH also assigned Z-scores above 3 to 12 alcohol dehydrogenases. In contrast, the discrete profile of 6LDH failed to assign any alcohol dehydrogenase sequence a Z-score above 3. Thus, the continuous profile with smoothing factor $B = 0.1$ is considerably more effective than the discrete profile in recognizing compatibility of structure with distantly related sequences folded the same way.

Comparing 3D structure compatibility search with 1D sequence homology search for common carp parvalbumin

We have shown above that in the structure compatibility searches for both sperm whale myoglobin and dogfish muscle lactate dehydrogenase, continuous profiles are more effective than their corresponding discrete profiles. Here we compare a 3D structure compatibility search with a sequence homology search (Gribskov et al., 1987) using common carp parvalbumin as the test protein. This protein has the further advantage that, unlike globins and dehydrogenases, it was not in the database of proteins used to derive the 3D-1D scores. Thus, the possibility of residual memory of sequence in 3D profiles is also tested.

A structure compatibility search was performed using the continuous 3D profile from the structure of common carp parvalbumin (4CPV) (Kumar et al., 1990). Similarly, a sequence homology search was performed using a sequence profile created from the sequence of 4CPV. The continuous profile identified 78 closely related sequences of parvalbumin, calmodulin, and troponin, whereas the sequence profile identified 67 of them. The continuous profile also detected 9 sequences of calcineurin, calretinin, and neurocalcin believed to have EF-hand fold, whereas the sequence profile failed to detect any of these sequences. Moreover, the continuous profile identified many myosin regulatory light chain sequences that have the same EF-hand motif as parvalbumin (Rayment et al., 1993). Figure 5 shows the distribution of the Z-scores for those myosin sequences both from the continuous profile compatibility search and from the sequence profile homology search. Sequences with Z-scores greater than 6 are generally folded in the same way as the structure represented by the profile. The continuous profile detected 56 myosin sequences with Z-score above 6; in contrast, none of the myosin sequences scored higher than 6 in the sequence profile. It is difficult to use sequence homology to infer structural similarity when sequence identity drops below 25% (Doolittle, 1986). The sequence identities of these myosin sequences detected by the continuous profile with that of the parvalbumin range from 16 to 30%, as shown in Figure 6. The majority of them are around 19–20%. This demonstrates that the continuous profile is sensitive in detecting distant structure–sequence relationships for which sequence identity is well below the 25% sequence level. Because the test is with a protein not used to determine 3D-1D scores, it also demonstrates that the effectiveness of the continuous 3D profile is not due to any residual memory of the sequence of the profiled structure.

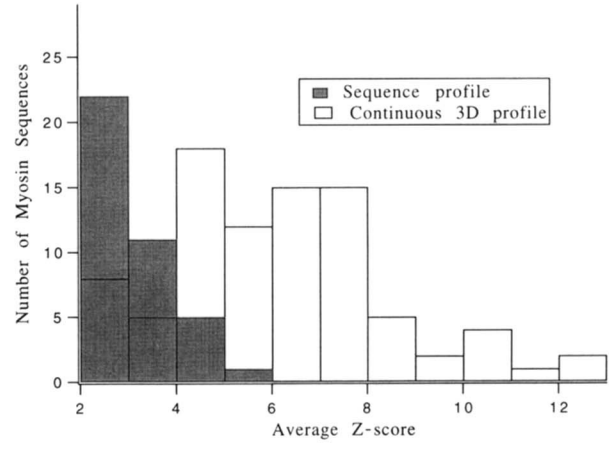

Fig. 5. Distribution of Z-scores for the myosin sequences identified by the sequence profile and the continuous profile of common carp parvalbumin (4CPV). The ordinate is the average Z-score for the myosin sequences. The abscissa is the number of myosin sequences at a given Z-score. A gap-opening penalty of 4.5 and gap-extension penalty of 0.05 were used for both the structure compatibility search and the sequence homology search. The continuous profile assigned Z-scores greater than 6 to many myosin sequences, in contrast to the sequences profile, which assigned no Z-scores higher than 6 to myosin sequences.

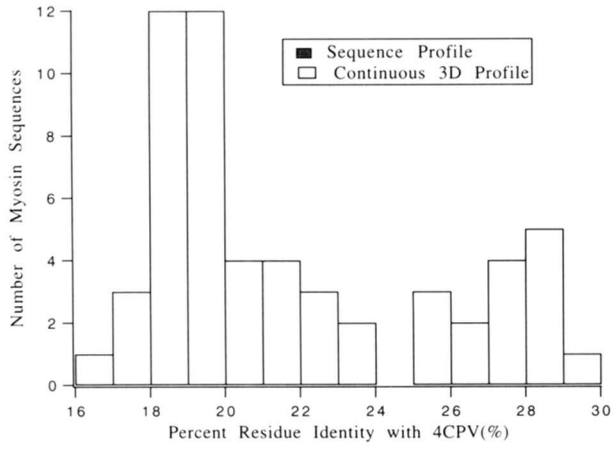

Fig. 6. Histogram of the sequence identities for all myosin sequences identified by the sequence and continuous profile of 4CPV with *Z*-score above 6. These sequences have sequence identities with the sequence of 4CPV ranging from 16 to 30%. Most of the sequences identified have sequence identities around 19–20%. This demonstrates that the continuous 3D profile is effective in recognizing distantly related sequences.

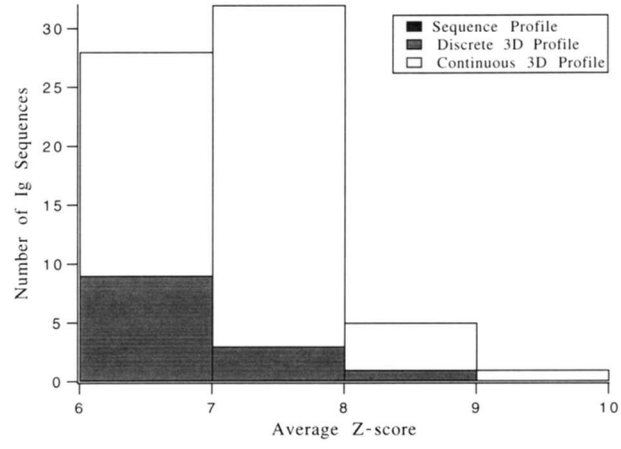

Fig. 7. Distribution of *Z*-scores for the immunoglobulin sequences identified by the sequence profile and the discrete and the continuous 3D profiles of Cu, Zn superoxide dismutase (2SOD). The ordinate is the average *Z*-score for the immunoglobulin sequences. The abscissa is the number of immunoglobulin sequences at a given *Z*-score. A gap-opening penalty of 4.5 and gap-extension penalty of 0.05 were used for both the structure compatibility search and the sequence homology search. The continuous 3D profile assigned *Z*-scores greater than 6 to 66 immunoglobulin sequences; in contrast, the discrete 3D profile assigned *Z*-scores greater than 6 to 13 immunoglobulin sequences, and the sequence profile assigned no *Z*-scores higher than 6 to immunoglobulin sequences.

The continuous profile compared with the discrete profile and the sequence profile for bovine superoxide dismutase

The greater sensitivity of the continuous profile in detecting distant relationships than either the sequence profile or the discrete profile is demonstrated further in the case of bovine superoxide dismutase. We created a continuous 3D profile and a discrete 3D profile from the structure of bovine Cu, Zn superoxide dismutase (2SOD) (Tainer et al., 1982) and also a sequence profile from the sequence of 2SOD. Each of these 3 profiles was used to score sequences in a database. The continuous, discrete, and sequence profiles all identified 51 closely related superoxide dismutase sequences. However, the continuous profile and the discrete profile were able to identify many immunoglobulin sequences that have a Greek-key fold similar to that of 2SOD (Richardson et al., 1976). Figure 7 shows the distribution of the *Z*-scores for those immunoglobulin sequences identified by these 3 profiles. The sequence profile failed to detect any immunoglobulin sequences with *Z*-score above 6. The discrete profile identified only 13 immunoglobulin sequences, as compared with 66 identified by the continuous profile. The sequence identities of these immunoglobulin sequences detected by the continuous profile with the sequence of 2SOD range from 11 to 26%, as shown in Figure 8. Their average sequence identity is around 16%. This shows that the continuous profile is more sensitive in detecting distant sequence–structure relationships than either the discrete profile or the sequence profile.

3D profile window plots

Sensitivity of 3D profile window plots to errors in the environmental variables

The 3D profile window plot (Lüthy et al., 1992) is a tool to assess the accuracy of a 3D protein structure, based on the compatibility of a structure with its own sequence. The average profile score for each 21-residue segment is plotted against sequence number. Segments scoring poorly have environments incompatible with the sequence and suggest errors in the structure of these segments. However, the details of a profile window plot prepared from a discrete 3D profile can vary significantly with the orientation of coordinates of the same structure, as explained below. We tested both the discrete and continuous profiles created from the same structure of myoglobin (1MBO) and rotated the coordinates randomly 9 times. A window plot for all the 9 coordinates was generated using discrete profiles, as shown in

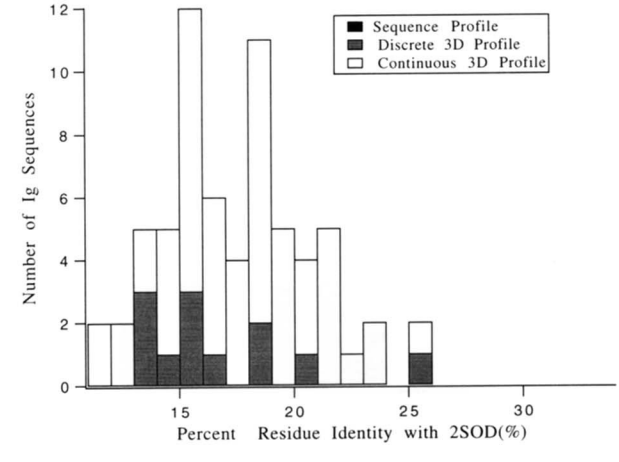

Fig. 8. Histogram of the sequence identities for all immunoglobulin sequences identified by the sequence profile and the discrete and the continuous 3D profiles of 2SOD with *Z*-score above 6. These immunoglobulin sequences identified by the continuous 3D profile have sequence identities with the sequence of 2SOD ranging from 11 to 26%. The average percent residue identity is 16%. This demonstrates that the continuous 3D profile is effective in recognizing distantly related sequences.

Figure 9A. The discrete profile showed significant variation in almost all regions. In contrast, the window plot created with the continuous profile (Fig. 9B) is almost unaffected by orientation. When examining the environmental classes of the discrete profiles for different orientations, it was found that the environmental classes changed for up to 9% of the residues, although the residues remain in exactly the same environment in the structure. The reason for the changes is the sampling error in estimating the area buried and fraction polar. Different orientations of coordinates yield slightly different values of area buried and fraction polar. Small as they may be, if they are near the border between 2 classes, this small difference is enough to change the classification of the residue environment. In contrast, for a continuous profile, because the residue preference varies smoothly with the area buried and fraction polar, it prevents an abrupt change of preference with a small change in orientation.

Detecting wrong folds and monitoring the progress of structure refinement

The continuous 3D profile can be used to detect wrong folds or an incorrect segment in an otherwise correct structure in the same way as the discrete profile (Lüthy et al., 1992). An incorrectly modeled region tends to have a score near or below zero in a profile window plot. Similarly, the 3D profile can also be used to monitor the progress of structure refinement. Generally, a well-refined structure has higher profile score than the unrefined structure. The improvement of the model at different regions can be shown from a profile window plot.

The detection of an incorrect segment of the model and the monitoring of the progress of refinement are illustrated for diphtheria toxin (DT) in Figure 10. Diphtheria toxin is a member of ADP-ribosylation toxins. The crystal structure of DT was solved initially at 2.5 Å by Choe et al. (1992) and refined to 2.0 Å by M.J. Bennett and D. Eisenberg ("The refined structure of dimeric diphtheria toxin at 2 Å resolution," ms. in prep.). In a preliminary trial model built into a 3.0-Å electron density map, much of the C-terminal receptor binding domain (R-domain) was at first reversed. However, it was rebuilt as the resolution and quality of the electron density map was improved by phase refinement and extension to 2.5 Å. Figure 10 shows the profile window plots of the trial model with reverse-traced R-domain, the 2.5-Å model, and the 2.0-Å-refined model. The profile scores in the R-domain (residues 380–518) are near or below zero, suggesting that this region was incorrect, as in fact it was. The R-domain in the 2.5-Å model has a significantly higher score, confirming that the trial model was incorrect. The scores in the profile window plot for the 2.0-Å-refined model are generally higher than those for the 2.5-Å model, reflecting the improved structure of the 2.0-Å-refined model. The places with the greatest improvement reflect an initial misregistration of the sequence with the electron density, which was corrected during rebuilding. The profile score for the 2.0-Å-refined model has one

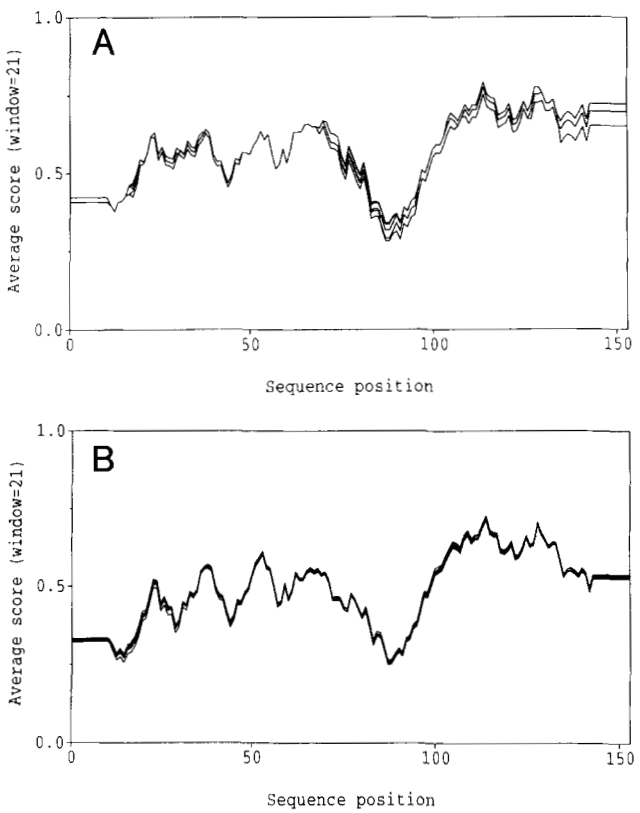

Fig. 9. A: Sensitivity of discrete 3D profiles, illustrated by a 3D profile window plot (Lüthy et al., 1992) for 1MBO with discrete 3D profiles created from coordinates rotated randomly 9 times. The vertical axis shows the average 3D-1D score for residues in a 21-residue sliding window, the center of which is at the sequence position indicated by the horizontal axis. Scores for the first and the last 9 sequence positions have no meaning due to the averaging. **B:** Stability of continuous 3D profiles, illustrated by 3D profile window plot for 1MBO with continuous 3D profiles created from coordinates rotated randomly 9 times, which are the same as the coordinates used in A. Notice that the fluctuation of profile scores in B is significantly less than that in A.

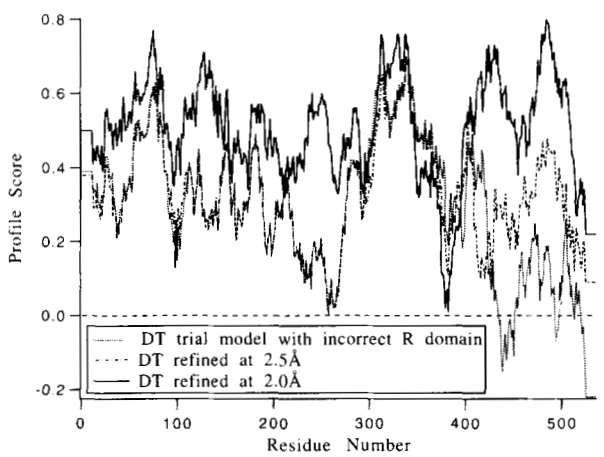

Fig. 10. Continuous 3D profile window plot for 3 models of diphtheria toxin. The abscissa is the position in the sequence of diphtheria toxin. The ordinate is the profile score averaged by a 21-residue window centered around that residue. The profile window plots for an initial trial model and for 2.5-Å and 2.0-Å models are represented by dotted, dashed, and solid lines, respectively. The profile scores in the C-terminal R-domain of the trial model are near or below zero, showing that this segment of the trial model was incorrect in this region. The profile score for the published 2.5-Å model with proper R-domain scored much higher than the corresponding region in the trial model. The profile scores are significantly higher in the 2.0-Å-refined model than in the 2.5-Å model. Notice the dip near residues 380–387 in the 2.0-Å-refined model corresponds to a hinge loop that is in a high energy state (Bennett & Eisenberg, in prep.).

noticeable dip near residues 380–387. These residues are in the hinge loop, which changes conformation when DT dimerizes by domain swapping, as discussed in detail by Bennett and Eisenberg (in prep.).

Discussion

In this paper we introduce continuous 3D profiles, in which the scores for protein side chains change smoothly as their environments change. These profiles have the same form and are used in the same way as the original discrete 3D profiles. However, compatibility searches based on continuous 3D profiles are more specific and selective than those of discrete 3D profiles. As illustrated by the tests presented here, the Z-scores for highly homologous sequences are almost always higher when using continuous 3D profiles in place of discrete 3D profiles. This suggests that the continuous 3D profile is more specific than the discrete 3D profile. Moreover, the continuous 3D profile is also more tolerant and accommodating for sequences that adopt the same fold as the profiled structure but share little sequence similarity. From the result shown in Figure 4, the continuous 3D profile detects more sequences of malate dehydrogenase with higher Z-scores than does the discrete 3D profile, and also assigns higher scores to the distantly related alcohol dehydrogenase sequences. This also demonstrates that a continuous 3D profile can be more selective than a discrete 3D profile.

Compatibility searches with profiles prepared from bovine Cu, Zn superoxide dismutase and common carp parvalbumin, using both the discrete and the continuous 3D profiles, further demonstrate that the continuous 3D profile is more sensitive to distant structure–sequence relationships than the discrete 3D profile. Moreover, the greater sensitivity of the continuous 3D profile over the sequence profile for both common carp parvalbumin and bovine superoxide dismutase demonstrates that the better specificity and selectivity of continuous 3D profiles is not the result of database bias. Neither of these 2 structures was included in the database of well-refined structures in obtaining the 3D-1D scoring table. In both these cases, the continuous profiles assigned significant profile scores to sequences that have only 15–20% sequence identity with the profile sequence. At the 15–20% level, structure similarity can be difficult to detect by sequence homology searches.

The continuous 3D profile is less sensitive to errors in the environmental variables than the discrete 3D profile, and yet it is also more sensitive to small differences in the model, for example at successive stages of refinement of an X-ray structure. This sensitivity is important not so much for identifying severe errors, but more for sensing slight changes of residue environment arising from slight changes of atomic position during refinement. The continuous preference function serves this purpose well. The test with a molecule having different orientations demonstrates that the continuous profile is virtually independent of the orientation, as it should be. The remaining slight variation of profile scores in the 3D profile window plot is due to the sampling error in the accessible surface area calculation.

The continuous 3D profile can be used to detect wrong folds or to monitor the progress of the structure improvement by a profile window plot. Because the continuous profile is less sensitive to measurement errors in the environmental variables, the differences between the profile window plots from 2 models can be confidently attributed to the differences in the quality of the models.

Notice that the reason we compute the residue preferences as a continuous function of environmental variables starting from data derived from only 6 environmental classes is because we lack enough data for a finer sampling. The advantage of the 6 environmental classes is that they recognize the essential characteristics of the environments, as specified by the variables area buried and fraction polar, and thus are able to bridge the gap between structure and sequence. A finer sampling of area buried and fraction polar can be used once a sufficient number of well-refined structures is available.

This method of representing residue preferences as a continuous function of environment can be easily extended to incorporate more continuous variables that characterize the environment. If more parameters were found to be useful determinants of protein stability, they can be incorporated into this scheme by adding other dimensions.

The representation of residue preference as a continuous function of residue environment (area buried and fraction polar) has an altogether new implication: the derivatives of residue preference with respect to the environmental variables, area buried and fraction polar, can be obtained analytically. This development offers the possibility of representing the residue preference as a continuous function of atomic positions, since area buried and fraction polar can be expressed as functions of atomic positions (Richmond, 1984; Wesson & Eisenberg, 1992; von Freyberg & Braun, 1993). Moreover, the derivatives of residue preference with respect to atomic positions can be obtained analytically. This opens the possibility of adding profile preferences to established refinement methods, by maximizing the residue preferences over the whole structure through a gradient-driven algorithm. The established refinement methods to which profile preferences could be added include X-ray and NMR refinements and energy refinements.

Acknowledgments

We thank Drs. James Bowie, Roland Lüthy, Laura Wesson, and Matthias Wilmanns for help and discussion and Dr. Ron Stenkamp for stimulating discussions on the orientation dependence of discrete 3D profiles. We also thank Drs. Seunghyon Choe and Melannie Bennett for their coordinates of diphtheria toxin structures. This work was supported by grants from DOE and NIH.

References

Bernstein FC, Koetzle TF, Williams GJB, Meyer EF Jr, Brice MD, Rodgers JR, Kennard O, Shimanouchi T, Tasumi M. 1977. The Protein Data Bank: A computer-based archival file for macromolecular structures. *J Mol Biol 112*:535–542.

Bowie JU, Lüthy R, Eisenberg D. 1991. A method to identify protein sequences that fold into a known three-dimensional structure. *Science 253*:164–170.

Choe S, Bennett MJ, Fujii G, Curmi PMG, Kantardjieff KA, Collier RJ, Eisenberg D. 1992. The crystal structure of diphtheria toxin. *Nature 357*:216–222.

Doolittle RF. 1986. *Of urfs and orfs: A primer on how to analyze derived amino acid sequences*. Mill Valley, California: University Science Books.

Fano R. 1961. *Transmission of information*. New York: Wiley.

Gribskov M, Lüthy R, Eisenberg D. 1990. Profile analysis. *Methods Enzymol 183*:146–159.

Gribskov M, McLachlan AD, Eisenberg D. 1987. Profile analysis: Detection of distantly related proteins. *Proc Natl Acad Sci USA 84*:4355–4358.

Kabsch W, Sander C. 1983. Dictionary of protein secondary structure: Pat-

tern recognition of hydrogen-bonded and geometric features. *Biopolymers 22*:2577–2637.

Kumar VD, Lee L, Edwards BFP. 1990. Refined crystal structure of calcium-liganded carp parvalbumin 4.25 at 1.5 Å resolution. *Biochemistry 29*:1404–1412.

Lüthy R, Bowie JU, Eisenberg D. 1992. Assessment of protein models with three-dimensional profiles. *Nature 356*:83–85.

Lüthy R, McLachlan A, Eisenberg D. 1991. Secondary structure-based profiles: Use of structure-conserving scoring tables in searching protein sequence databases for structural similarities. *Proteins Struct Funct Genet 10*:229–239.

Needleman SB, Wunsch CD. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol 48*:443–453.

Rao ST, Rossmann MG. 1973. Comparison of super-secondary structures in proteins. *J Mol Biol 76*:241–256.

Rayment I, Rypniewski WR, Schmidt-Bäse K, Smith R, Tomchick DR, Benning MM, Winkelmann DA, Wesenberg G, Holden HM. 1993. Three-dimensional structure of myosin subfragment-1: A molecular motor. *Science 261*:50–58.

Richardson JS, Richardson DC, Thomas KA, Silverton EW, Davies DR. 1976. Similarity of three-dimensional structure between the immunoglobulin domain and the copper, zinc superoxide dismutase subunit. *J Mol Biol 102*:221–235.

Richmond TJ. 1984. Solvent accessible surface area and excluded volume in proteins. *J Mol Biol 178*:63–89.

Smith TF, Waterman MS. 1981. Identification of common molecular subsequences. *J Mol Biol 147*:195–197.

Tainer JA, Getzoff ED, Beem KM, Richardson JS, Richardson DC. 1982. Determination and analysis of the 2 Å structure of copper, zinc superoxide dismutase. *J Mol Biol 160*:181–217.

von Freyberg B, Braun W. 1993. Minimization of empirical energy functions in proteins including hydrophobic surface area effects. *J Comput Chem 14*:510–521.

Wesson L, Eisenberg D. 1992. Atomic solvation parameters applied to molecular dynamics of proteins in solution. *Protein Sci 2*:227–235.